

Sparse Matrix Inversion with Scaled Lasso

Tingni Sun and Cun-Hui Zhang
Rutgers University

Address: Department of Statistics and Biostatistics, Hill Center, Busch Campus, Rutgers University, Piscataway, New Jersey 08854, U.S.A.

E-mail addresses: tingni@stat.rutgers.edu, czhang@stat.rutgers.edu

Abstract

We propose a new method of learning a sparse nonnegative-definite target matrix. Our primary example of the target matrix is the inverse of a population covariance matrix or correlation matrix. The algorithm first estimates each column of the matrix by scaled Lasso, a joint estimation of regression coefficients and noise level, and then adjusts the matrix estimator to be symmetric. The procedure is efficient in the sense that the penalty level of the scaled Lasso for each column is completely determined by the data via convex minimization, without using cross-validation. We prove that this method guarantees the fastest proven rate of convergence in the spectrum norm under conditions of weaker form than those in the existing analyses of other ℓ_1 algorithms, and has faster guaranteed rate of convergence when the ratio of the ℓ_1 and spectrum norms of the target inverse matrix diverges to infinity. A simulation study also demonstrates the competitive performance of the proposed estimator.

1 Introduction

We consider the estimation of the matrix inversion Θ^* satisfying $\bar{\Sigma}\Theta^* \approx I$, given a data matrix $\bar{\Sigma}$. When $\bar{\Sigma}$ is a sample covariance matrix, our problem is the estimation of the inverse of the corresponding population covariance matrix. The inverse covariance matrix is also called precision matrix or concentration matrix. With the dramatic advances in technology, the number of covariates is of greater order than the sample size n in many statistical and engineering applications. In this case, the sample covariance matrix is always singular and thus it is difficult to compute the precision matrix. In such cases, a certain type of sparsity condition is required for proper estimation the precision matrix and for theoretical investigation of the estimation problem. In this paper, we will impose for simplicity an ℓ_0 (maximum degree) sparsity condition on the target inverse matrix Θ^* .

Many approaches have been proposed to estimate the sparse inverse matrix in the high dimensional setting. The ℓ_1 penalization is one of the most popular methods. Lasso-type methods, or convex minimization algorithms with the ℓ_1 penalty on all entries of Θ^* , have been discussed by Banerjee, El Ghaoui and d'Aspremont

(2008), Friedman, Hastie and Tibshirani (2008) and more, and by Yuan and Lin (2007) with ℓ_1 penalization on the off-diagonal matrix only. This is referred to as the graphical Lasso (GLasso) due to the connection of the precision matrix to Gaussian Markov graphical models. In this GLasso framework, Rothman, Bickel, Levina and Zhu (2008) proved the convergence rate $\{((p+s)/n) \log p\}^{1/2}$ in the Frobenius norm and $\{(s/n) \log p\}^{1/2}$ in the spectrum norm, where s is the number of nonzero entries in the off-diagonal matrix. Ravikumar, Wainwright, Raskutti, and Yu (2008) provided sufficient conditions for model selection consistency of this ℓ_1 -regularized MLE. Lam and Fan (2009) studied on a general penalty function and achieved a sharper bound of order $\{(s/n) \log p\}^{1/2}$ under the Frobenius norm for the ℓ_1 penalty. Since the spectrum norm can be controlled via the Frobenius norm, this provides a sufficient condition $(s/n) \log p \rightarrow 0$ for the convergence under the spectrum norm to the unknown precision matrix. This is a very strong condition since s is of the order dp for banded precision matrices, where d is the matrix degree, i.e. the largest number of nonzero entries in the columns.

Some recent work suggests a weaker sufficient condition with the matrix degree. Yuan (2010) estimated each column of the inverse matrix by Dantzig selector and then seek a symmetric matrix close to the column estimation. When ℓ_1 norm of the precision matrix is bounded, this method can achieve a convergence rate of order $d\{(\log p)/n\}^{1/2}$ based on several matrix norms. The CLIME estimator, introduced by Cai, Liu and Luo (2011), has the same order of convergence rate, which uses the plug-in method with Dantzig selector to estimate each column, but followed by a simpler symmetrization step. They also require the boundedness of the ℓ_1 norm of the unknown. In Yang and Kolaczyk (2010), the Lasso is applied to estimate the columns of the target matrix under the assumption of equal diagonal, and the estimation error is studied in the Frobenius norm for $p = n^\nu$. This column-by-column idea reduces a graphical model to a regression model. It was first introduced in Meinshausen and Bühlmann (2006) for identifying nonzero variables in a graphical model, called neighborhood selection.

In this paper, we propose to apply the scaled Lasso (Sun and Zhang, 2011) column-by-column to estimate a precision matrix in the high dimensional setting. Based on the connection of precision matrix to linear regression by the block inversion formula, we construct a column estimator with the scaled Lasso, a joint estimator for the regression coefficients and noise level. Since we only need the sample covariance matrix in our procedure, this estimator could be extended to generate an approximate inverse of a nonnegative data matrix in a general setting. This scaled Lasso algorithm provides a fully specified map from the space of nonnegative-definite matrices to the space of symmetric matrices. For each column, the penalty level of the scaled Lasso is determined by data via convex minimization, without using cross-validation.

We study theoretical properties of the proposed estimator for a precision matrix under a normality assumption. More precisely, we assume that the data matrix is the sample covariance matrix $\bar{\Sigma} = \mathbf{X}'\mathbf{X}/n$, where the rows of \mathbf{X} are iid $N(0, \Sigma^*)$. Under

conditions on the spectrum norm and degree of the inverse of Σ^* , we prove that the proposed estimator guarantees the rate of convergence of order $d\{(\log p)/n\}^{1/2}$ in the spectrum norm. The conditions are weaker than those in the existing analyses of other ℓ_1 algorithms, which typically require the boundedness of the ℓ_1 norm. When the ℓ_1 norm of the target matrix diverges to infinity, the analysis of the proposed estimator guarantees a faster convergence rate than that of the existing literature. We state this main result of the paper in the following theorem.

Theorem 1 *Let $\hat{\Theta}$ be the scaled Lasso estimator, defined in (5), (6) and (7) below, based on n iid observations from $N(0, \Sigma^*)$. Let ρ_* and ρ^* be the smallest and largest eigenvalues of correlation matrix of Σ^* , Θ^* be the inverse of Σ^* and $d = \max_i \#\{j : \Theta_{ij}^* \neq 0\}$ be the maximum degree of Θ^* . Suppose that $d\sqrt{(\log p)/n} \rightarrow 0$, the diagonal entries of the target matrix Θ^* are uniformly bounded, ρ_* is bounded from 0 and $(\rho^*/\rho_*)\{(d/n)\log p\}^{1/2} < a$ for a small fixed a . Then, the spectrum norm of the estimation error $\hat{\Theta} - \Theta^*$ is bounded by*

$$\|\hat{\Theta} - \Theta^*\|_2 = O_P(d\sqrt{(\log p)/n}).$$

The convergence of the proposed scaled Lasso estimator under the sharper spectrum norm condition on Θ^* , instead of the stronger bounded ℓ_1 condition, is not entirely technical. It is a direct consequence of the faster convergence rate of the scaled Lasso estimator of the noise level in linear regression. To the best of our knowledge, it is unclear if other ℓ_1 algorithms also achieve this fast convergence rate, either for the estimation of the noise level in linear regression or for the estimation of a precision matrix under the spectrum norm. However, it is still possible that this difference between the scaled Lasso and other methods is due to potentially coarser specification of the penalty level in other algorithms (e.g. cross validation) or a less accurate error bound in other analyses.

The rest of the paper is organized as follows. In Section 2, we present the estimation for the inversion of a nonnegative definite matrix via the scaled Lasso. In Section 3, we study error bounds of the proposed estimator for precision matrix. Simulation studies are presented in Section 4. In Section 5, we discuss oracle inequalities for the scaled Lasso with unnormalized predictors and the estimation of inverse correlation matrix. Section 6 includes all the proofs.

We use the following notation throughout the paper. For a vector $\mathbf{v} = (v_1, \dots, v_p)$, $\|\mathbf{v}\|_q = (\sum_j |v_j|^q)^{1/q}$ is the ℓ_q norm with the special $\|\mathbf{v}\| = \|\mathbf{v}\|_2$ and the usual extensions $\|\mathbf{v}\|_\infty = \max_j |v_j|$ and $\|\mathbf{v}\|_0 = \#\{j : v_j \neq 0\}$. For matrices \mathbf{M} , $\mathbf{M}_{j,*}$ is the j -th column of \mathbf{M} , $\mathbf{M}_{A,B}$ represents the submatrix of \mathbf{M} with rows in A and columns in B , $\|\mathbf{M}\|_q = \sup_{\|\mathbf{v}\|_q=1} \|\mathbf{M}\mathbf{v}\|_q$ is the ℓ_q matrix norm. In particular, $\|\cdot\|_2$ is the spectrum norm for symmetric matrices. Moreover, we denote the set $\{j\}$ by j and denote the set $\{1, \dots, p\} \setminus \{j\}$ by $-j$ in the subscripts.

2 Matrix inversion via scaled Lasso

Let $\bar{\Sigma}$ be a nonnegative-definite data matrix and Θ^* be a positive-definite target matrix with $\bar{\Sigma}\Theta^* \approx \mathbf{I}$. In this section, we describe the relationship between positive-definite matrix inversion and linear regression and propose an estimator for Θ^* via scaled Lasso, a joint convex minimization for the estimation of regression coefficients and noise level.

We use scaled Lasso to estimate Θ^* column by column. Define $\sigma_j > 0$ and $\beta \in \mathbb{R}^{p \times p}$ by

$$\sigma_j^2 = (\Theta_{jj}^*)^{-1}, \quad \beta_{*,j} = -\Theta_{*,j}^* \sigma_j^2 = -\Theta_{*,j}^* (\Theta_{jj}^*)^{-1}. \quad (1)$$

In the matrix form, we have the following relationship

$$\text{diag} \Theta^* = \text{diag}(\sigma_j^{-2}, j = 1, \dots, p), \quad \Theta^* = -\beta(\text{diag} \Theta^*). \quad (2)$$

Let $\Sigma^* = (\Theta^*)^{-1}$. Since $(\partial/\partial \mathbf{b}_{-j}) \mathbf{b}' \Sigma^* \mathbf{b} = 2 \Sigma_{-,j}^* \mathbf{b} = 0$ at $\mathbf{b} = \beta_{*,j}$, one may estimate the j -th column of β by minimizing the ℓ_1 penalized quadratic loss. In order to shrink the estimation coefficients on the same scale, we adjust the penalty function with a normalizing factor, which leads to the ℓ_1 penalized quadratic loss as follows,

$$\mathbf{b}' \bar{\Sigma} \mathbf{b} / 2 + \lambda \sum_{k=1}^p \bar{\Sigma}_{kk}^{-1/2} |b_k|$$

subject to $b_j = -1$. This is actually the Lasso for a linear regression model with normalized predictors. In practice, we first normalize the predictors by the weights $\bar{\Sigma}_{kk}^{-1/2}$ ($k \neq j$) and then the minimization problem can be solved by algorithms for the Lasso estimation. This is similar to Yuan (2010) and Cai, Liu and Luo (2011) who used the Dantzig selector to estimate each column. However, one still needs to choose a penalty level λ and to estimate σ_j to recover Θ^* via (2). A solution to resolve these two issues is the scaled Lasso (Sun and Zhang, 2011):

$$\{\hat{\beta}_{*,j}, \hat{\sigma}_j\} = \arg \min_{\mathbf{b}, \sigma} \left\{ \frac{\mathbf{b}' \bar{\Sigma} \mathbf{b}}{2\sigma} + \frac{\sigma}{2} + \lambda_0 \sum_{k=1}^p \bar{\Sigma}_{kk}^{-1/2} |b_k| : b_j = -1 \right\} \quad (3)$$

where $\lambda_0 = A \sqrt{2(\log p^2/\epsilon)/n}$ with a fixed $A > 1$. The scaled Lasso (3) is a solution of joint convex minimization over $\{\mathbf{b}, \sigma\}$ due to the convexity in (\mathbf{b}, σ) (Huber, 2009; Antoniadis, 2010). Since $\beta' \Sigma^* \beta = (\text{diag} \Theta^*)^{-1} \Theta^* (\text{diag} \Theta^*)^{-1}$,

$$\text{diag}(\beta' \Sigma^* \beta) = (\text{diag} \Theta^*)^{-1} = \text{diag}(\sigma_j^2, j = 1, \dots, p).$$

Thus, (3) is expected to yield consistent estimates of σ_j .

Sun and Zhang (2011) provided an iterative algorithm to compute the scaled Lasso estimator (3). We rewrite the algorithm in the form of matrices. For each $j \in$

$\{1, \dots, p\}$, the Lasso path is given by the estimates $\widehat{\beta}_{-j,j}(\lambda)$ satisfying the following KKT conditions, for all $k \neq j$,

$$\begin{cases} \overline{\Sigma}_{kk}^{-1/2} \overline{\Sigma}_{k,*} \widehat{\beta}_{*,j}(\lambda) = -\lambda \text{sgn}(\widehat{\beta}_{k,j}(\lambda)), & \widehat{\beta}_{k,j} \neq 0, \\ \overline{\Sigma}_{kk}^{-1/2} \overline{\Sigma}_{k,*} \widehat{\beta}_{*,j}(\lambda) \in \lambda[-1, 1], & \widehat{\beta}_{k,j} = 0, \end{cases} \quad (4)$$

where $\widehat{\beta}_{jj}(\lambda) = -1$. Based on the Lasso path $\widehat{\beta}_{*,j}(\lambda)$, the scaled Lasso estimator $\{\widehat{\beta}_{*,j}, \widehat{\sigma}_j\}$ is computed iteratively by

$$\widehat{\sigma}_j^2 \leftarrow \widehat{\beta}_{*,j}' \overline{\Sigma} \widehat{\beta}_{*,j}, \quad \lambda \leftarrow \widehat{\sigma}_j \lambda_0, \quad \widehat{\beta}_{*,j} \leftarrow \widehat{\beta}_{*,j}(\lambda). \quad (5)$$

Here the penalty level of the Lasso is determined by the data without using cross-validation. We then simply take advantage of the relationship (2) and compute the coefficients and noise levels by the scaled Lasso for each column

$$\text{diag} \widetilde{\Theta} = \text{diag}(\widehat{\sigma}_j^{-2}, j = 1, \dots, p), \quad \widetilde{\Theta} = -\widehat{\beta}(\text{diag} \widetilde{\Theta}). \quad (6)$$

It is noticed that a good estimator for Θ^* should be a symmetric matrix. However, the estimator $\widetilde{\Theta}$ does not have to be symmetric. We improve this estimator by using a symmetrization step as in Yuan (2010),

$$\widehat{\Theta} = \arg \min_{\mathbf{M}: \mathbf{M}^T = \mathbf{M}} \|\mathbf{M} - \widetilde{\Theta}\|_1, \quad (7)$$

which can be solved by linear programming. Alternatively, semidefinite programming, which is somewhat more expensive computationally, can be used to produce a nonnegative definite $\widehat{\Theta}$ in (7). According to the definition, the new estimator $\widehat{\Theta}$ has the same ℓ_1 error rate as $\widetilde{\Theta}$. A nice property for symmetric matrix is that the spectrum norm is bounded by the ℓ_1 matrix norm. The ℓ_1 matrix norm can be given more explicitly as the maximum ℓ_1 norm of the columns, while the ℓ_∞ matrix norm is the maximum ℓ_1 norm of the rows. Hence, for any symmetric matrix, the ℓ_1 matrix norm is equivalent to the ℓ_∞ matrix norm, so the spectrum norm can be bounded by either of them. Since both our estimator and the target matrix are symmetric, the error bound based on the spectrum norm could be studied by bounding the ℓ_1 error, as typically done in the existing literature. We will discuss these error bounds in Section 3.

To sum up, we propose to estimate the matrix inversion by (5), (6) and (7). The iterative algorithm (5) computes the regression coefficients and noise level based on a Lasso path determined by (4). Then (6) translates the resulting estimators of (5) to column estimators and thus a preliminary matrix estimator is constructed. Finally, the symmetrization step (7) produces a symmetric estimate for our target matrix.

3 Error bounds for precision matrix

In this section, we study the error $\widehat{\Theta} - \Theta^*$ for the inverse of a covariance matrix, which is our primary example of the target matrix. From now on, we suppose that the data matrix is the sample covariance matrix $\bar{\Sigma} = \mathbf{X}'\mathbf{X}/n$, where the rows of \mathbf{X} are iid $N(0, \Sigma^*)$, and the target matrix is $\Theta^* = (\Sigma^*)^{-1}$.

Let ρ_* and ρ^* be the smallest and the largest eigenvalues of the correlation matrix $(\text{diag}\Sigma^*)^{-1/2}\Sigma^*(\text{diag}\Sigma^*)^{-1/2}$. Define $S_j = \{i \neq j : \Theta_{i,j}^* \neq 0\}$ and the degree of the matrix

$$d = \deg(\Theta^*) = \max_j |S_j| + 1.$$

The following theorem gives the convergence rate based on the ℓ_1 matrix norm (ℓ_∞ matrix norm) and spectrum norm.

Theorem 2 *Let $\epsilon \in (0, 1/4)$ and $\lambda_0 = A\{(2/n)\log(p^2/\epsilon)\}^{1/2}$ with $A > 1$. Suppose that $\{d(\log p)/n\}^{1/2}\rho^*/\rho_* < a$ for a small fixed a . Then with probability greater than $1 - 4\epsilon$,*

$$\begin{aligned} \|\widehat{\Theta} - \Theta^*\|_2 &\leq \|\widehat{\Theta} - \Theta^*\|_1 = \|\widehat{\Theta} - \Theta^*\|_\infty \\ &\leq C_1\lambda_0^2d\|\Theta^*\|_1\rho_*^{-1} + C_2\lambda_0d\max_k \Theta_{kk}^*\rho_*^{-1} \end{aligned} \quad (8)$$

where C_1 and C_2 are constants depending on $\{A, a\}$ only.

Since the entries of Θ^* are bounded by the maximum of the diagonal, the ℓ_1 matrix norm $\|\Theta^*\|_1$ is of the same order as the matrix degree d . Thus, the inequality (8) provides a convergence rate of the order $d\lambda_0$ for either the ℓ_1 matrix norm or the spectrum norm under the conditions $d\{(\log p)/n\}^{1/2} \rightarrow 0$, $\rho_*^{-1} = O(1)$ and $\max(\Theta^*)_{kk} = O(1)$. The first condition is the main sparsity condition, and the other two are actually conditions on the ℓ_2 norm of the target matrix. To achieve the same convergence rate, Yuan (2010) and Cai, Liu and Luo (2011) both imposed the condition $d\{(\log p)/n\}^{1/2} \rightarrow 0$ and the boundedness of the ℓ_1 norm of the unknown. We replace the ℓ_1 condition by the weaker boundedness of the spectrum norm of the unknown. The spectrum norm condition on the unknown is not only weaker, but also natural for the convergence in spectrum norm. The extra condition $\{d(\log p)/n\}^{1/2}\rho^*/\rho_* < a$ here is not strong. Under the conditions $d\{(\log p)/n\}^{1/2} \rightarrow 0$ and $\rho_*^{-1} = O(1)$, the extra condition only requires $\rho^*/d^{1/2}$ to be small and it allows ρ^* to diverge to infinity.

This sharper error bound in the spectrum norm is a consequence of using the scaled Lasso estimator (3). Sun and Zhang (2011) gave a convergence rate of order λ_0^2d for the scaled Lasso estimation of the noise levels σ_j . With this faster rate of convergence, the estimation error in the diagonal is no longer the main term and thus the condition of the bounded ℓ_1 norm of Θ^* can be weakened.

The consistency of the scaled Lasso estimation for the noise level is based on the ℓ_1 error bound for the regression coefficients. Oracle inequalities for the ℓ_1 error of the Lasso have been studied with various conditions, including the restricted isometry condition (Candes and Tao, 2007), the compatibility condition (van de Geer, 2007) and the sign-restricted cone invertibility factor (Ye and Zhang, 2010) among others. Sun and Zhang (2011) extended these oracle inequalities for the scaled Lasso. Here we use the version under the condition of ℓ_1 sign-restricted cone invertibility factor (SCIF)

$$SCIF_1(\xi, S; \Sigma) = \inf \left\{ \frac{|S| \cdot \|\Sigma \mathbf{u}\|_\infty}{\|\mathbf{u}\|_1} : \mathbf{u} \in \mathcal{C}_-(\xi, S) \right\} > 0, \quad (9)$$

with the cone $\mathcal{C}(\xi, S) = \{\mathbf{u} \in R^{p-1} : \|\mathbf{u}_{S^c}\|_1 \leq \xi \|\mathbf{u}_S\|_1\}$ and the sign-restricted cone $\mathcal{C}_-(\xi, S) = \{\mathbf{u} \in \mathcal{C}(\xi, S) : u_j \Sigma_{j,*} \mathbf{u} \leq 0, \forall j \notin S\}$. It is proved that, conditional on $\mathbf{X}_{*, -j}$,

$$\left| \frac{\hat{\sigma}_j}{\sigma_j} - 1 \right| = O_p(1) |S_j| \lambda_0^2, \quad \|\hat{\beta}_{-j,j} - \beta_{-j,j}\|_1 / \sigma_j = O_p(1) |S_j| \lambda_0, \quad (10)$$

under the condition that $SCIF_1(\xi, S_j; \bar{\Sigma}_{-j})$ is bounded away from 0. This is guaranteed by the conditions of Theorem 2. The error bound of ℓ_1 matrix norm then follows from (10).

4 Numerical study

In this section, we compare the proposed matrix estimator based on scaled Lasso with graphical Lasso and CLIME (Cai, Liu and Luo, 2011). Three models are considered. The first two models are the same as model 1 and model 2 in Cai, Liu and Luo (2011). Model 2 was also studied in Rothman et al. (2008).

- Model 1: $\Theta_{ij} = 0.6^{|i-j|}$.
- Model 2: Let $\Theta = \mathbf{B} + \delta \mathbf{I}$, where each off-diagonal entry in \mathbf{B} is generated independently and equals to 0.5 with probability 0.1 or 0 with probability 0.9. δ is chosen such that the condition number of Θ^* is p . Finally, we rescale the matrix Θ^* to the unit in diagonal.
- Model 3: The diagonal of the target matrix has unequal values. $\Theta = D^{1/2} \Omega D^{1/2}$, where $\Omega_{ij} = 0.6^{|i-j|}$ and D is a diagonal matrix with diagonal elements $d_{ii} = (4i + p - 5) / \{5(p - 1)\}$.

For each model, we generate a training sample of size 100 from a multivariate normal distribution with mean zero and covariance matrix $\Sigma = \Theta^{-1}$ and an independent sample of size 100 from the same distribution for validating the tuning parameter λ for

the graphical Lasso and CLIME. The GLasso and CLIME estimators are computed based on training data with various λ 's and we choose λ by minimizing likelihood loss $\{\text{trace}(\widehat{\Sigma\Theta}) - \log \det(\widehat{\Theta})\}$ on the validation sample. The proposed scaled Lasso estimator is computed based on the training sample alone with the penalty level $\lambda_0 = \{(\log p)/n\}^{1/2}$. Consider 6 different dimensions $p = 30, 60, 90, 150, 300, 1000$ and replicate 100 times for each case. The CLIME estimators for $p = 300$ and $p = 1000$ are not computed due to the computational costs.

Table 1 presents the mean and standard deviation of estimation errors based on 100 replications. The estimation error is measured by several matrix norms: spectrum norm, matrix ℓ_1 norm and Frobenius norm. We can see that scaled Lasso estimator, labelled as SLasso, outperforms the graphical Lasso (GLasso) in all cases, while it has a comparable performance with the CLIME.

Table 1: Estimation errors under various matrix norms of scaled Lasso, GLasso and CLIME for three models.

Model 1									
p	Spectrum norm			Matrix ℓ_1 norm			Frobenius norm		
	SLasso	GLasso	CLIME	SLasso	GLasso	CLIME	SLasso	GLasso	CLIME
30	2.41(0.08)	2.49(0.14)	2.29(0.21)	2.93(0.11)	3.09(0.11)	2.92(0.17)	4.09(0.12)	4.24(0.26)	3.80(0.36)
60	2.61(0.05)	2.94(0.05)	2.68(0.10)	3.10(0.09)	3.55(0.07)	3.27(0.09)	6.16(0.10)	7.15(0.15)	6.32(0.28)
90	2.67(0.05)	3.07(0.03)	2.87(0.09)	3.19(0.08)	3.72(0.06)	3.42(0.07)	7.73(0.11)	9.25(0.12)	8.42(0.31)
150	2.74(0.04)	3.19(0.02)	3.05(0.04)	3.28(0.08)	3.88(0.06)	3.55(0.06)	10.22(0.13)	12.55(0.09)	11.68(0.20)
300	2.80(0.03)	3.29(0.01)	NA	3.38(0.07)	4.06(0.05)	NA	14.77(0.11)	18.44(0.09)	NA
1000	2.87(0.03)	3.39(0.00)	NA	3.52(0.06)	4.44(0.07)	NA	27.59(0.12)	35.11(0.06)	NA
Model 2									
p	Spectrum norm			Matrix ℓ_1 norm			Frobenius norm		
	SLasso	GLasso	CLIME	SLasso	GLasso	CLIME	SLasso	GLasso	CLIME
30	0.75(0.08)	0.82(0.07)	0.81(0.09)	1.32(0.16)	1.49(0.15)	1.45(0.18)	1.90(0.10)	1.84(0.09)	1.87(0.11)
60	1.07(0.05)	1.15(0.06)	1.19(0.08)	1.97(0.16)	2.21(0.12)	2.20(0.23)	3.31(0.08)	3.18(0.13)	3.42(0.09)
90	1.49(0.04)	1.54(0.05)	1.61(0.04)	2.63(0.16)	2.89(0.16)	2.90(0.17)	4.50(0.06)	4.40(0.11)	4.65(0.08)
150	1.98(0.03)	2.02(0.05)	2.06(0.03)	3.31(0.17)	3.60(0.15)	3.65(0.19)	6.02(0.05)	6.19(0.16)	6.33(0.08)
300	2.85(0.02)	2.89(0.02)	NA	4.50(0.14)	4.92(0.17)	NA	9.35(0.05)	9.79(0.05)	NA
1000	5.35(0.01)	5.52(0.01)	NA	7.30(0.21)	7.98(0.15)	NA	18.34(0.05)	20.81(0.02)	NA
Model 3									
p	Spectrum norm			Matrix ℓ_1 norm			Frobenius norm		
	SLasso	GLasso	CLIME	SLasso	GLasso	CLIME	SLasso	GLasso	CLIME
30	1.75(0.10)	2.08(0.10)	1.63(0.19)	2.24(0.14)	2.59(0.10)	2.17(0.20)	2.52(0.10)	2.91(0.16)	2.37(0.25)
60	2.09(0.08)	2.63(0.04)	2.10(0.10)	2.58(0.13)	3.10(0.05)	2.65(0.14)	3.81(0.09)	4.84(0.08)	3.98(0.13)
90	2.24(0.07)	2.84(0.03)	2.38(0.18)	2.72(0.12)	3.30(0.06)	2.91(0.12)	4.79(0.08)	6.25(0.08)	5.37(0.37)
150	2.40(0.06)	3.06(0.02)	2.76(0.05)	2.89(0.11)	3.45(0.04)	3.18(0.09)	6.35(0.09)	8.43(0.07)	7.75(0.08)
300	2.54(0.05)	3.26(0.01)	NA	3.05(0.10)	3.58(0.03)	NA	9.20(0.09)	12.41(0.04)	NA
1000	2.68(0.05)	3.47(0.01)	NA	3.26(0.09)	3.73(0.03)	NA	17.2(0.09)	23.55(0.02)	NA

5 More results

5.1 Oracle inequalities for scaled Lasso estimator

In the proof of the theoretical results for the proposed estimator, we use oracle inequalities for the estimation error associated with a linear model without normalizing the predictors. In the discussion section, we describe this aspect of our results. Consider a linear model as follows,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I}_n).$$

Let $\mathbf{\Sigma} = \mathbf{X}'\mathbf{X}/n$, $\mathbf{D} = \text{diag}\mathbf{\Sigma}$, $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{D}^{1/2}$ and $\widetilde{\mathbf{\Sigma}} = \mathbf{D}^{-1/2}\mathbf{\Sigma}\mathbf{D}^{-1/2}$. In order to penalize the coefficients on the same scale, we use a weighted ℓ_1 norm of the coefficients as the penalty function. Consider the estimator

$$\{\widehat{\boldsymbol{\beta}}, \widehat{\sigma}\} = \arg \min_{\mathbf{b}, \sigma} \left\{ \frac{\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 \|\mathbf{D}^{1/2}\mathbf{b}\|_1 \right\}. \quad (11)$$

This is actually the scaled Lasso as we use in matrix estimation in Section 2. It is equivalent to the estimation based on normalized predictors:

$$\{\widehat{\boldsymbol{\alpha}}, \widehat{\sigma}\} = \arg \min_{\mathbf{a}, \sigma} \left\{ \frac{\|\mathbf{y} - \widetilde{\mathbf{X}}\mathbf{a}\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 \|\mathbf{a}\|_1 \right\} \quad (12)$$

with $\widehat{\boldsymbol{\beta}} = \mathbf{D}^{-1/2}\widehat{\boldsymbol{\alpha}}$.

The following theorem gives the oracle inequalities for the estimation of regression coefficients and noise level.

Theorem 3 *Let $\{\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\sigma}\}$ be as in (11) and (12), $\sigma^* = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2/n^{1/2}$, $S = \{k : \beta_k \neq 0\}$, $z^* = \|\widetilde{\mathbf{X}}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/n\|_\infty/\sigma^*$ and $\xi > 1$.
(i) In the event $z^* \leq (1 + \tau^+)^{-1/2}\lambda_0(\xi - 1)/(\xi + 1)$,*

$$\frac{1}{1 + \tau^+} \leq \left(\frac{\widehat{\sigma}}{\sigma^*}\right)^2 \leq \frac{1}{1 - \tau^-}, \quad \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_1 \leq \frac{(1 + \xi)\tau^- \sigma^*}{2\xi\lambda_0(1 - \tau^-)^{1/2}}, \quad (13)$$

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq \frac{(1 + \xi)\tau^- \sigma^*}{2\xi\lambda_0(1 - \tau^-)^{1/2} \min_k D_{kk}^{1/2}}, \quad (14)$$

where $\tau^- = \phi_1(\xi)\lambda_0^2|S|/SCIF_1(\xi, S; \widetilde{\mathbf{\Sigma}})$ and $\tau^+ = \phi_2(\xi)\lambda_0^2|S|/SCIF_1(\xi, S; \widetilde{\mathbf{\Sigma}})$ with constants $\xi > 1$, $\phi_1(\xi) = 4\xi^2/(1 + \xi)^2$ and $\phi_2(\xi) = 4\xi(\xi - 1)/(1 + \xi)^2$.

(ii) Let $\lambda_0 \geq \{(2/n)\log(p/\epsilon)\}^{1/2}(\xi + 1)/\{(\xi - 1)(1 - \tau_-)\}$. For $n - 2 > \log(p/\epsilon) \rightarrow \infty$,

$$P\{z^* \leq (1 - \tau_-)\lambda_0(\xi - 1)/(\xi + 1)\} \geq 1 - (1 + o(1))\epsilon/\sqrt{\pi \log(p/\epsilon)}.$$

Theorem 3 is an immediate extension from the oracle inequalities for the scaled Lasso in Sun and Zhang (2011). With an extra condition that $\mathbf{x}'_k \mathbf{x}_k/n (k = 1, \dots, p)$ are uniformly bounded from zero, the estimators have the same convergence rate as that for a regression model with normalized predictors. The error rates (10) follows from Theorem 3 and are used to prove the convergence rate of matrix estimation.

5.2 Error bounds for inverse correlation matrix

Given the data matrix $\overline{\mathbf{\Sigma}} = \mathbf{X}'\mathbf{X}/n$, we are also interested in estimating the inverse correlation matrix

$$\mathbf{\Omega}^* = \{\mathbf{D}^{-1/2}\mathbf{\Sigma}^*\mathbf{D}^{-1/2}\}^{-1} = \mathbf{D}^{1/2}(\mathbf{\Sigma}^*)^{-1}\mathbf{D}^{1/2}.$$

where $\mathbf{D} = \text{diag}(\boldsymbol{\Sigma}^*)$. When constructing the matrix estimator $\tilde{\boldsymbol{\Theta}}$, we solve the linear regression problem with normalized predictors. If the estimator $\hat{\boldsymbol{\beta}}$ in (6) is replaced by $\hat{\boldsymbol{\alpha}}$ as in (12), the resulting matrix is an estimator for $\mathbf{D}^{1/2}(\boldsymbol{\Sigma}^*)^{-1}$. Thus, the inverse correlation matrix is estimated by

$$\begin{aligned}\tilde{\boldsymbol{\Omega}} &= -\hat{\boldsymbol{\alpha}} \text{diag}(\hat{\sigma}_j^{-2} \bar{\boldsymbol{\Sigma}}_{jj}^{1/2}, j = 1, \dots, p), \\ \hat{\tilde{\boldsymbol{\Omega}}} &= \arg \min_{\mathbf{M}: \mathbf{M}^T = \mathbf{M}} \|\mathbf{M} - \tilde{\boldsymbol{\Omega}}\|_1.\end{aligned}\tag{15}$$

The error bounds of this estimator are well established as follows.

Theorem 4 *Let $\epsilon \in (0, 1/4)$, $\lambda_0 = A\{(2/n) \log(p^2/\epsilon)\}^{1/2}$ with $A > 1$ and $d\lambda_0 \rightarrow 0$. Suppose that $\{d(\log p)/n\}^{1/2} \rho^*/\rho_* < a$ for a small fixed a . Then with probability greater than $1 - 4\epsilon$,*

$$\begin{aligned}\|\hat{\tilde{\boldsymbol{\Omega}}} - \boldsymbol{\Omega}^*\|_2 &\leq \|\hat{\tilde{\boldsymbol{\Omega}}} - \boldsymbol{\Omega}^*\|_1 = \|\hat{\tilde{\boldsymbol{\Omega}}} - \boldsymbol{\Omega}^*\|_\infty \\ &\leq C_1 \lambda_0^2 d \|\boldsymbol{\Omega}\|_1 \|\boldsymbol{\Omega}\|_2 + C_2 \lambda_0 \|\boldsymbol{\Omega}\|_1 + C_3 \lambda_0 d \|\boldsymbol{\Omega}\|_2 \max_{jj} \Omega_{jj}^{1/2}\end{aligned}\tag{16}$$

where C_1 , C_2 and C_3 are constants depending on $\{A, a\}$ only.

Theorem 4 shows that the error bounds are also of the order $d\lambda_0$ under proper conditions. The crucial condition here is still the boundedness of the spectrum norm of the unknown matrix.

6 Proofs

In this section, we provide the proofs of Theorem 3, Theorem 2 and Theorem 4. Theorem 1 is a brief version of Theorem 2, so we omit the proof.

Proof of Theorem 3. The inequalities (13) are parallel to Theorem 2 in Sun and Zhang (2011). The only difference is that here we use the ℓ_1 bound under the condition of the sign-restricted cone invertibility factor (SCIF). Since $\hat{\boldsymbol{\beta}} = \mathbf{D}^{-1/2} \hat{\boldsymbol{\alpha}}$, (14) follows from the second inequality in (13).

Proof of Theorem 2. Let $\xi > (A + 1)/(A - 1)$, $(\sigma_j^*)^2 = \boldsymbol{\beta}'_{*,j} \bar{\boldsymbol{\Sigma}} \boldsymbol{\beta}_{*,j}$, $z_{(j),k} = \bar{\boldsymbol{\Sigma}}_{kk}^{-1/2} |\bar{\boldsymbol{\Sigma}}_{k,*} \boldsymbol{\beta}_{*,j}| / \sigma_j^*$ and $z_{(j)}^* = \max_{k \neq j} z_{(j),k}$. By Theorem 4, in the event $z_{(j)}^* \leq (1 + \tau_{(j)}^+)^{-1/2} \lambda_0 (\xi - 1) / (\xi + 1)$,

$$\frac{1}{1 + \tau_{(j)}^+} \leq \left(\frac{\hat{\sigma}_j}{\sigma_j^*} \right)^2 \leq \frac{1}{1 - \tau_{(j)}^-}, \quad \sum_{k \neq j} \bar{\boldsymbol{\Sigma}}_{kk}^{1/2} |\hat{\beta}_{k,j} - \beta_{k,j}| \leq \frac{(1 + \xi) \tau_{(j)}^- \sigma_j^*}{2\xi \lambda_0 (1 - \tau_{(j)}^-)^{1/2}},\tag{17}$$

where $\tau_{(j)}^- = \phi_1(\xi) \lambda_0^2 |S_j| / \text{SCIF}_1(\xi, S_j; \tilde{\boldsymbol{\Sigma}}_{-j})$ and $\tau_{(j)}^+ = \phi_2(\xi) \lambda_0^2 |S_j| / \text{SCIF}_1(\xi, S_j; \tilde{\boldsymbol{\Sigma}}_{-j})$.

We first derive some probabilistic bounds for some useful quantities. Since $\bar{\Sigma} = \mathbf{X}'\mathbf{X}/n$ and the rows of \mathbf{X} follow a multivariate normal distribution with covariance matrix Σ^* , we have $\sigma_j^* = \|\mathbf{x}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_{-j,j}\|/\sqrt{n}$ and $z_{(j),k} = \tilde{\mathbf{x}}_k(\mathbf{x}_j - \mathbf{X}_{-j}\boldsymbol{\beta}_{-j,j})/\sigma_j^*$. Thus, $n(\sigma_j^*/\sigma_j)^2$ follows a χ^2 distribution with n degrees of freedom and thus

$$P\{|(\sigma_j^*/\sigma_j)^2 - 1| > \sqrt{(8/n)\log(2p/\epsilon)}\} \leq \epsilon/p. \quad (18)$$

Also, we have that $z_{(j),k}/\{(1 - z_{(j),k}^2)/(n-1)\}^{1/2}$ follows a t -distribution with $n-1$ degrees of freedom. We use Lemma 1 in Sun and Zhang (2011) with $m = n-1$ and $t^2 = \log(p^2/\epsilon) > 2$ to obtain

$$P\{|z_{(j),k}| > \sqrt{2\log(p^2/\epsilon)/n}\} \leq (1 + \epsilon_{n-1})(\epsilon/p^2)/\sqrt{\pi\log(p^2/\epsilon)}.$$

Thus,

$$P\{\max_j |z_{(j)}^*| > \sqrt{2\log(p^2/\epsilon)/n}\} \leq \epsilon, \quad (19)$$

i.e. the events $z_{(j)}^* \leq (1 + \tau_{(j)}^+)^{-1/2} \lambda_0(\xi-1)/(\xi+1)$ ($j = 1, \dots, p$) occur with probability greater than $1 - \epsilon$. Since $\bar{\Sigma}_{kk} \sim \Sigma_{kk}^* \chi_n^2/n$, we have

$$P\{|\bar{\Sigma}_{kk}/\Sigma_{kk}^* - 1| > \sqrt{(8/n)\log(2p/\epsilon)}\} \leq \epsilon/p. \quad (20)$$

So there exists a small ζ , such that $\max |\bar{\Sigma}_{kk}/\Sigma_{kk}^* - 1| < \zeta$ holds for all k with probability greater than $1 - \epsilon$.

Now we need to bound $SCIF_1(\xi, S_j; \tilde{\Sigma}_{-j})$, for all j , with probability greater than $1 - \epsilon$ under the given conditions, where $S_j = \{i \neq j : \beta_{i,j} \neq 0\}$. Let $\mathbf{Z} = \mathbf{X}(\text{diag}\Sigma^*)^{-1/2}$. We discuss the bounds for $SCIF_1$ within the event $\max |\bar{\Sigma}_{kk}/\Sigma_{kk}^* - 1| < \zeta$.

For $(|A|, |B|, \|\mathbf{u}\|, \|\mathbf{v}\|_r) = (\lceil a \rceil, \lceil b \rceil, 1, 1)$ with $A \cap B = \emptyset$, we define

$$\delta_a^\pm = \delta_a^\pm(\mathbf{X}) = \max_{A, \mathbf{u}} \left\{ \pm \left(\|\mathbf{X}'_A \mathbf{X}_A \mathbf{u}/n\| - 1 \right) \right\}, \quad \theta_{a,b}^{(2)} = \theta_{a,b}^{(2)}(\mathbf{X}) = \max_{A, B, \mathbf{u}, \mathbf{v}} \mathbf{v}' \mathbf{X}'_A \mathbf{X}_B \mathbf{u}/n.$$

For any subset $T \subset \{1, \dots, p\}$, we have

$$\theta_{a,b}^{(2)} \geq \theta_{a,b}^{(2)}(\mathbf{X}_T), \quad \delta_a^\pm \geq \delta_a^\pm(\mathbf{X}_T), \quad \theta_{a,b}^{(2)} \leq (1 + \delta_a^+)^{1/2} (1 + \delta_b^+)^{1/2} \leq 1 + \delta_{a \vee b}^+. \quad (21)$$

By Proposition 2(i) in Zhang and Huang (2008), we have

$$P\left\{(1-c)^2 \rho_* \leq 1 - \delta_m^-(\mathbf{Z}) \leq 1 + \delta_m^+(\mathbf{Z}) \leq (1+c)^2 \rho^*\right\} \geq 1 - \epsilon, \quad (22)$$

where $c = \sqrt{m/n} + \sqrt{(2m/n)\log(2p/\epsilon)}(1 + o(1))$. We also have

$$1 + \delta_a^+(\tilde{\mathbf{X}}) \leq \max(\Sigma_{kk}^*/\bar{\Sigma}_{kk})(1 + \delta_a^+(\mathbf{Z})) = (1 + \delta_a^+(\mathbf{Z}))/ (1 - \zeta),$$

$$1 - \delta_a^-(\widetilde{\mathbf{X}}) \geq \min(\Sigma_{kk}^*/\overline{\Sigma}_{kk})(1 - \delta_a^-(\mathbf{Z})) = (1 - \delta_a^-(\mathbf{Z}))/(1 + \zeta). \quad (23)$$

Let $k_j = |S_j|$. It follows from the shifting inequality in Ye and Zhang (2010) with $\ell \geq d$ that

$$\begin{aligned} SCIF_1(\xi, S_j; \widetilde{\Sigma}_{-j}) &\geq \frac{1}{1 + \xi} (1 - \delta_{k_j + \ell}^-(\widetilde{\mathbf{X}}_{-j}) - \xi \sqrt{\frac{k_j}{4\ell}} \theta_{4\ell, k_j + \ell}^{(2)}(\widetilde{\mathbf{X}}_{-j})) \\ &\geq \frac{1}{1 + \xi} \{1 - \delta_{4\ell}^-(\widetilde{\mathbf{X}}) - \xi \sqrt{\frac{d}{4\ell}} (1 + \delta_{4\ell}^+(\widetilde{\mathbf{X}}))\} \\ &\geq \frac{1}{1 + \xi} \left\{ \frac{1 - \delta_{4\ell}^-(\mathbf{Z})}{1 + \zeta} - \xi \sqrt{\frac{d}{4\ell}} \frac{1 + \delta_{4\ell}^+(\mathbf{Z})}{1 - \zeta} \right\} \end{aligned}$$

The second and the third inequalities follow from (21) and (23), respectively. Let $m = 4\ell$ in (22) with $\ell = d(\xi\rho^*/\rho_*)^2 > d$. Then

$$SCIF_1(\xi, S_j; \widetilde{\Sigma}_{-j}) \geq \frac{\rho_*}{1 + \xi} \left\{ \frac{(1 - c)^2}{1 + \zeta} - \frac{(1 + c)^2}{2(1 - \zeta)} \right\}.$$

Under the condition $(\rho^*/\rho_*)\{(d/n)\log p\}^{1/2} < a$ for a small fixed a , c is also very small. Thus, with probability greater than $1 - \epsilon$, $SCIF_1(\xi, S_j; \widetilde{\Sigma}_{-j})$ are bounded by $C\rho_*$ for all j , where C is a constant only depending on $\{\xi, \zeta, a\}$.

Now we are ready to bound ℓ_1 of the column of $\widetilde{\Theta} - \Theta$ by (17), (18), (19), (20) and the uniform bound for $SCIF_1$. The following inequalities hold with probability greater than $1 - 4\epsilon$:

$$\begin{aligned} \|\widetilde{\Theta}_{\cdot j} - \Theta_{\cdot j}^*\|_1 &\leq |\widetilde{\Theta}_{jj} - \Theta_{jj}^*| + \|\widetilde{\Theta}_{-j,j} - \Theta_{-j,j}^*\|_1 \\ &\leq \|\Theta_{\cdot j}\|_1 \cdot \left| \left(\frac{\widehat{\sigma}_j}{\sigma_j} \right)^{-2} - 1 \right| + \left(\frac{\widehat{\sigma}_j}{\sigma_j^*} \right)^{-2} \left(\frac{\sigma_j^*}{\sigma_j} \right)^{-1} \sigma_j^{-1} \frac{\|\widehat{\beta}_{-j,j} - \beta_{-j,j}\|_1}{\sigma_j^*} \\ &\leq \|\Theta_{\cdot j}\|_1 \frac{C'_1 \lambda_0^2 |S_j|}{\rho_*} + \frac{C'_2 \lambda_0 |S_j|}{\sigma_j (\min_k \Sigma_{kk}^*)^{1/2} \rho_*}. \end{aligned}$$

The first two inequalities just use some simple algebra, while the last one put (17), (18), (19), (20) and the uniform bound for $SCIF_1$ together. The constants C'_1 and C'_2 only depend on $\{A, a\}$. Therefore, the ℓ_1 error of the matrix estimator $\widetilde{\Theta}$ is bounded by

$$\|\widetilde{\Theta} - \Theta^*\|_1 \leq C'_3 \lambda_0^2 d \|\Theta^*\|_1 \rho_*^{-1} + C'_4 \lambda_0 d \max \Theta_{kk}^* \rho_*^{-1}.$$

Then the upper bound for $\|\widehat{\Theta} - \Theta\|_1$ follows from the triangle inequality and the definition of $\widehat{\Theta}$, since $\|\widehat{\Theta} - \widetilde{\Theta}\|_1 \leq \|\Theta^* - \widetilde{\Theta}\|_1$.

For any matrix \mathbf{M} and vector \mathbf{u}, \mathbf{v} , we have

$$\mathbf{u}' \mathbf{M} \mathbf{v} = \sum_{i,j} M_{ij} u_i v_j \leq \left(\sum_{i,j} M_{ij} u_i^2 \sum_{i,j} M_{ij} v_j^2 \right)^{1/2} \leq (\|\mathbf{M}\|_\infty \cdot \|\mathbf{M}\|_1)^{1/2} \|\mathbf{u}\| \cdot \|\mathbf{v}\|.$$

So $\|\mathbf{M}\|_2^2 \leq \|\mathbf{M}\|_\infty \cdot \|\mathbf{M}\|_1$. For the symmetric matrix $\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$, we have

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_2 \leq \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_\infty \leq \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1.$$

The desired error bounds based the spectrum norm then follows. \square

Proof of Theorem 4. The inequalities (18), (19), (20) and the uniform bound for $SCIF_1$ still hold. We first bound the ℓ_1 norm of one column of $\widetilde{\boldsymbol{\Omega}} - \boldsymbol{\Omega}$ as follows

$$\begin{aligned} & \|\widetilde{\boldsymbol{\Omega}}_{\cdot j} - \boldsymbol{\Omega}_{\cdot j}^*\|_1 \\ & \leq |\widetilde{\boldsymbol{\Omega}}_{jj} - \boldsymbol{\Omega}_{jj}^*| + \|\widetilde{\boldsymbol{\Omega}}_{-j,j} - \boldsymbol{\Omega}_{-j,j}^*\|_1 \\ & \leq |\widehat{\sigma}_j^{-2} \overline{\Sigma}_{jj} - \sigma_j^{-2} D_{jj}| + \widehat{\sigma}_j^{-2} \overline{\Sigma}_{jj}^{1/2} \|\widehat{\boldsymbol{\alpha}}_{-j,j} - \boldsymbol{\alpha}_{-j,j}\|_1 + \|\boldsymbol{\alpha}_{-j,j}\|_1 \cdot |\widehat{\sigma}_j^{-2} \overline{\Sigma}_{jj}^{1/2} - \sigma_j^{-2} D_{jj}^{1/2}| \\ & \leq |\widehat{\sigma}_j^{-2} \overline{\Sigma}_{jj} - \overline{\Sigma}_{jj}^{1/2} \sigma_j^{-2} D_{jj}^{1/2}| + \|\boldsymbol{\Omega}_{\cdot j}\|_1 \cdot \left| \frac{\widehat{\sigma}_j^{-2} \overline{\Sigma}_{jj}^{1/2}}{\sigma_j^{-2} D_{jj}^{1/2}} - 1 \right| + \widehat{\sigma}_j^{-2} \overline{\Sigma}_{jj}^{1/2} \|\widehat{\boldsymbol{\alpha}}_{-j,j} - \boldsymbol{\alpha}_{-j,j}\|_1 \\ & \leq C'_1 \Omega_{jj} \lambda_0^2 d / \rho_* + C'_2 \|\boldsymbol{\Omega}_{\cdot j}\|_1 (\lambda_0^2 d / \rho_* + C'_3 \sqrt{(\log p)/n}) + C'_4 \Omega_{jj}^{1/2} \lambda_0 d / \rho_*, \end{aligned}$$

where the constants only depend on $\{A, a\}$. Thus, taking the maximum for both sides gives the error bounds for the matrix estimator $\widetilde{\boldsymbol{\Omega}}$ under the ℓ_1 matrix norm. The rest proof for error bounds of $\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}^*$ under various norms are the same as that in the proof of Theorem 2.

References

- [1] ANTONIADIS, A. (2010). Comments on: ℓ_1 -penalization for mixture regression models. *Test.* **19**(2) 257-258.
- [2] BANERJEE, O., EL GHAOU, L. and D'ASPREMONT, A. (2008) Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research.* **9** 485-516.
- [3] CAI, T., LIU, W. and LUO, X. (2011) A Constrained ℓ_1 Minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association.* **106** 594-607.
- [4] CANDÈS, E. and TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35** 2313-2404.
- [5] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics.* **9** 432-441.
- [6] HUBER, P.J. and RONCHETTI, E.M. (2009). Robust statistics. *Wiley, 2nd edition.* pp. 172-175.

- [7] LAM, C. and FAN, J. (2009). Sparsistency and Rates of Convergence in Large Covariance Matrices Estimation. *Ann. Statist.* **37** 4254-4278.
- [8] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436-1462.
- [9] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2008) Model selection in Gaussian graphical models: High-dimensional consistency of l1-regularized MLE. *In Advances in Neural Information Processing Systems (NIPS)* 21.
- [10] ROTHMAN, A.J., BICKEL, P.J., LEVINA, E., and ZHU, J. (2008). Sparse Permutation Invariant Covariance Estimation. *Electronic Journal of Statistics.* **2** 494-515.
- [11] SUN, T. and ZHANG, C.-H. (2011). Scaled sparse linear regression. *arXiv: 1104.4595v1*.
- [12] VAN DE GEER, S. (2007). The deterministic Lasso. In *JSM proceedings*, American Statistical Association.
- [13] YANG, S. and Kolaczyk, E. D. (2010) Target detection via network filtering. *IEEE Transactions on Information Theory.* **56** (5) 2502-2515.
- [14] YE, F. and ZHANG, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research.* **11** 3481-3502.
- [15] YUAN, M. (2010). Sparse Inverse Covariance Matrix Estimation via Linear Programming. *Journal of Machine Learning Research.* **11** 2261-2286.
- [16] YUAN, M. and LIN, Y. (2007). Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika.* **94** (1) 19-35.
- [17] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567-1594.